VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

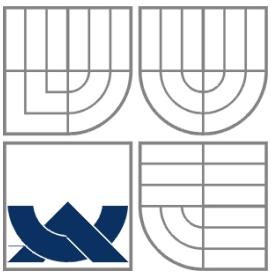
EXTRAKCE METADAT Z VĚDECKÝCH ČLÁNKŮ

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

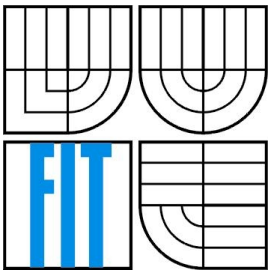
AUTOR PRÁCE
AUTHOR

PETR VÁCHA

BRNO 2009



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

EXTRAKCE METADAT Z VĚDECKÝCH ČLÁNKŮ

METADATA EXTRACTION FROM SCIENTIFIC PAPERS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

PETR VÁCHA

VEDOUCÍ PRÁCE
SUPERVISOR

Doc. RNDr. PAVEL SMRŽ, Ph.D.

BRNO 2009

Abstrakt

Práce porovnává dostupné vědecké vyhledávače se skriptem pro extrakci metadat z vědeckých prací vyvinutým Tomášem Lokajem na FIT VUT v Brně. Výsledky potvrzují nedostatky v extrakci metadat. Tato bakalářská práce zároveň představuje ucelený návod na porovnání různých informací.

Abstract

Work compares accessible scientific locater and program for extraction metadata from scientific papers created by Tomáš Lokaj on FIT BUT. Results affirm imperfections in extraction metadata. This bachelor thesis introduce integral manual for comparing various informations.

Klíčová slova

CiteSeerX, Google Scholar, extrakce metadat, vědecké články

Keywords

CiteSeerX, Google Scholar, metadata extraction, scientific papers

Citace

Petr Vácha: Extrakce metadat z vědeckých článků, bakalářská práce, Brno, FIT VUT v Brně, 2009

Extrakce metadat z vědeckých článků

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením doc. RNDr. Pavla Smrže, Ph.D. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……
Petr Vácha
26.5.2009

Poděkování

Rád bych poděkoval vedoucímu mé bakalářské práce, panu doc. RNDr. Pavlovi Smržovi, Ph.D., za jeho odbornou pomoc a vstřícné jednání. Dále Robertovi Kalmárovi za jeho ochotu poskytnutí svého upraveného programu pro detekci kvality PDF převodů.

Obsah

1 Úvod

Tato bakalářská práce představuje jednu z možností jak získávat metadata z internetu a porovnávat je s jiným, lokálně dostupným systémem. Práce se zaměřuje na pojem metadat v oblasti vědeckých publikací. Představuje přední dostupné webové systémy pro zpracování takovýchto prací a porovnává je s nástrojem pro extrakci metadat z vědeckých publikací vyvíjeným na FIT VUT v Brně[1]. Vhodně ukázat postup získávání a zpracování metadat z vědeckých vyhledávačů je jednou z priorit bakalářské práce. Na zvolených systémech jsou porovnávány jejich možnosti a nastavení. Dále by si práce měla zakládat na názorném zobrazení extrahovaných údajů. Představit navržený a autorem implementovaný program. V závěru porovnat kvalitu jednotlivých systémů a zhodnotit dosažené výsledky statistického zpracování. Tímto dát zpětnou vazbu fakultně vyvíjenému systému. Pro větší názornost jsou mnohé postupy a výsledky graficky zobrazeny. Autor v závěru sebekriticky? hodnotí vlastní systém a navrhuje další úpravy a směr, kterým by se měl systém dále vyvíjet.

1.1 Přehled kapitol

Jednotlivé kapitoly představují podrobný návod, jak od nejzákladnějších údajů postupovat k získání potřebných informací. Následující kapitola číslo 2 uvádí základní informace nutné k porozumění celé práce. Pro práci důležitá webová rozhraní jsou zde graficky znázorněna. Kapitola číslo 3, Analýza zvolených systémů, popisuje jednotlivé systémy do větší hloubky. K porovnání vhodná metadata z webových systémů jsou zde názorně zobrazena. Skript pro lokální zpracování je taktéž důsledně popsán. Neopomenut není ani jeho XML[2] výstup. V následující kapitole 4 srovnáme dostupné informace, zaměříme se na získání dat z HTML[3] tagů a popíšeme návrh systému pro zpracování těchto údajů. Poslední 5. kapitola představuje získané a vyhodnocené údaje.

1 Fakulta Informačních Technologií, Vysoké Učení Technické v Brně

2 eXtensible Markup Language, česky rozšiřitelný značkovací jazyk

3 HyperText Markup Language

2 Teoretická část

Tato velká kapitola čtenáře seznamuje se základními aspekty nutnými k porozumění problematiky rozebírané v dalších kapitolách.

2.1 Metadata

Metadata nebo také metainformace, jsou strukturovaná data o datech. Popisují data, které nemusí být vhodně strukturovaná. Mohou popisovat důležité části textu, obsah obrázku, videa, článků, ale i kvalitu objektu v zadaném měřítku. Jsou to data, která pomáhají pochopit a interpretovat význam popisovaných dat v konkrétním kontextu. Zahrnují nejen informace o datech samých, tedy o tom, co znamenají, v jakém jsou formátu, odkud pocházejí nebo jakých mohou nabývat hodnot, ale také informace o jejich vzájemných interakcí a vzájemné ovlivnitelnosti. [1]. Metadata jsou tedy ideální pro řešení vyhledávání na internetu. Pěkným a jednoduchým příkladem metadat jsou validně a slušně napsané webové stránky. Dříve se většina internetových vyhledávačů primárně řídila meta tagy webových stránek. Typicky tagy *description*, *keywords*, *Content-Language*, *Content-Type*, které slouží částečně jak pro vyhledávače, tak i pro webové prohlížeče. Jsou jasně definované a mají jasně viditelné opodstatnění. Pomocí těchto tagů si mohly/mohou webové vyhledávače jednoduše najít relevantní informace, které by měly popisovat obsah celé stránky. Jen pro doplnění – webové vyhledávače se tagy *description* a *keywords* již přestávají řídit nebo mají přísné hodnotící metriky. Důvodem je častá nesměrodatnost obsahu tagů vůči informacím na stránce.

2.2 BibTeX

BibTeX představuje nástroj a formát obvykle používaný ve spojení s LaTeXovými dokumenty. Byl vytvořen kolem roku 1985 Orenem Patashnikem. BibTeX umožňuje jednoduché a hlavně oddělené prezentování bibliografických informací. Většinou ve zvláštním souboru s koncovkou *.bib*. Bibliografické položky jsou pro jednotlivé typy citací povinné nebo volitelné. Mezi nejčastěji uvedené položky patří: *author, booktitle, chapter, editor, journal, title, url, volume, year, pages*. Položka *author* popisuje nejen jednoho autoru, ale může obsahovat i více jmen. Mezi typy záznamů patří: *article, book, booklet, conference, inbook, incollection, inproceeding, mastersthesis, misc, phdthesis, proceedings* aj.[2, 3, 4]

Příklad BibTeXového formátu:

```
@bachelorsthesis{vacha09,
  author    =     {Petr Vacha},
  title     =     {Metadata extraction from scientific papers},
  year      =     {2009},
  address   =     {Brno, CZ},
}
```

2.3 Vědecký článek

Vědecký článek se od různých neodborných článků, když ještě pomineme obsah, jakožto myšlenku či nový poznatek k předání, především liší v jeho striktních typografických a strukturálních konvencích, které si většinou nárokují pravidla odborných časopisů. Jde především o požadovaný jazyk publikace, použitý editor, dodržování struktury a formy jednotlivých částí. Struktura příspěvku by měla většinou obsahovat následující body:

- název příspěvku (title)
- jméno a adresa autora, popř. autorů[1]
- abstrakt (abstract)
- klíčová slova (keywords)
- úvod (introduction)
- jádro článku: materiály a metody, výsledky, diskuse
- poděkování (acknowledgement)
- použitou literaturu (references)

Všechny předchozí body budeme považovat za důležité metadata. Konvence pro psaní jmen autorů a jejich emailových či univerzitních adres jsou dnes bohužel nejednotné. Jména autorů, tak mohou být psané částečně iniciálami či celými jmény. Jednou z úvodních částí, která sděluje hlavní rysy vědeckého článku, je abstrakt. Pokud autor či autoři článku cítí potřebu poděkovat například za zkontrolování po odborné nebo po jazykové stránce nebo je článek financován grantovou agenturou nebo mají jiné důvody, sdělí svoje poděkování v toto odstavci. Mezi zásadní odstavce patří i použitá literatura, která je takřka zákonnou a věrohodnou podmínkou. Rozlišují se především dvě formy citování - vancouverské (dle pořadí výskytu v textu) nebo harvardské (abecední). [5]

1 Adresou je myšlena adresa působiště autora, ale i jeho emailová adresa.

2.4 Vyhledávače pro vědu

Ačkoliv se to nemusí zdát, tak v posledních letech vzniklo velké množství vyhledávačů či citačních databází pro vyhledávání věděckých článků. Vyhledávače slouží především k snadnému šíření vědeckých poznatků, k efektivnějšímu hledání a některé jako je třeba hodně komplexní databáze Web of Science, slouží k určení citačního indexu. Citační index odborného článku je seznam, který obsahuje publikace, které citovaly daný odborný článek. Pomocí citačního indexu se určuje impakt faktor neboli faktor vlivu článku. Ten by měl být nejdůležitějším ukazatelem vědecké úrovně publikace. Takovýto systém, který umí vyhodnotit citační index určité publikace, si musí evidovat vzájemné vztahy mezi publikacemi, jen díky extrahovaným referencím. [6]

Vyhledávače můžeme dělit na profitující, neboli komerční, a nekomerční, ty jsou zřejmě jen dotované nebo sponzorované. Od tohoto rozdělení se odvíjí způsob použití a nabízené služby. U komerčních systémů jsou dost často ke stažení vědecké práce v celém znění. Nutnou podmínkou je buď platit jednorázově za daný dokument, nebo být stálým členem, popřípadě pracovat v organizaci, typicky univerzita, která má členství v daném systému. Nekomerční projekty taktéž občas nabízejí odkaz na stažení dané publikace. Za zmínku stojí, že mnohdy i placené dokumenty lze dohledat zcela zdarma ke stažení v nekomerčních vyhledávačích. (Jsou to ovšem pouze odkazy na jiné servery.) [7]

Vyhledávače nebo akademické databáze jsou většinou zaměřené na určitý vědní obor nebo nějaký definovaný vědní rozsah např. astronomie, geofyzika, fyzika, jako má ADS[1].

Samozřejmě jsou i takové, které nejsou takto úzce zaměřené na konkrétní části vědy. Do této skupiny patří giganti: Google Scholar Beta, Scirus, Science research, WorldWideScience.org, komerční databáze Springer a mnoho dalších. Rozdělení takovýchto sytémů může být i na lokální nebo nadnárodní, kde dominuje anglický jazyk. U komerčních a u Googlu Scholar se můžeme setkat s přívětivou multilingvální politikou.

2.4.1 Google Scholar Beta

Google Scholar[2] Beta, dále už jen GS, je vědecký vyhledávač provozovaný společnosti Google. Spuštěn byl v listopadu roku 2004 jako odpověď na vyhledávač Scirus. GS na svých stránkách[8] tvrdí, že nabízí k vyhledání „informace z mnoha oborů a zdrojů: recenzované články, dizertační práce, knihy, abstrakty a články od akademických nakladatelství, odborných společností, archivů preprintů a dalších odborných organizací.“ GS se při hodnocení článku a následných dotazech

1 Astrophysics Data System, http://adsabs.harvard.edu

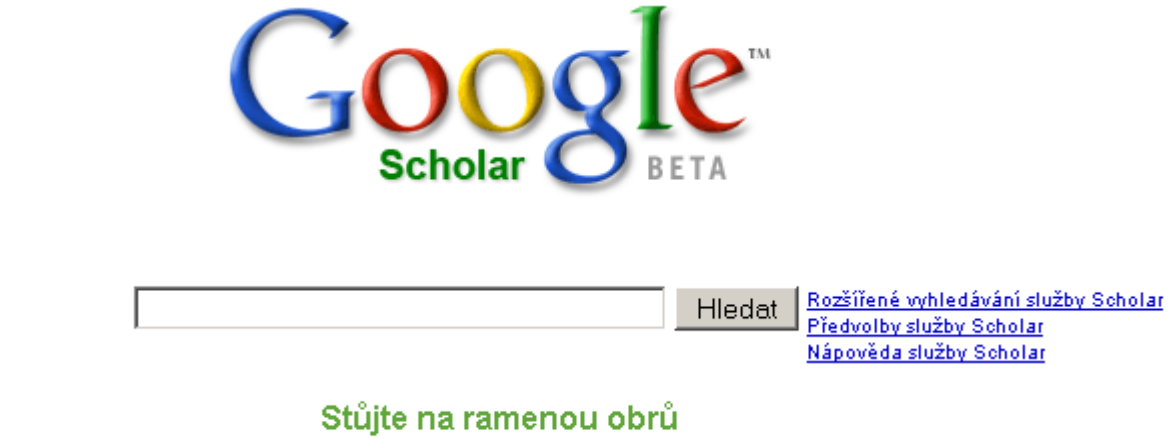
2 http://scholar.google.cz/

rozumně snaží brát v potaz celý obsah článku, jméno autora, publikační zdroje článku a kolikrát byl článek citován v jiné vědecké literatuře. Po dotazu se v drtivé většině nejrelevantnější výsledky objevují na první stránce. [8]

Zanedlouho po spuštění GS doplnil pokročilé možnosti vyhledávání a svým tempem rozvoje se stává silnou konkurencí placeným bibliografickým databázím. Stejně tak jako Scirus, GS nezveřejňuje všechny své zdroje odkud čerpá cenné informace.[9] Na rozdíl od ostatních mezinárodních vyhledávačů, je GS, jak je to u společnosti Google zvykem, kompletně počeštěn.

GS nabízí univerzitám či jiným vědecky zaměřeným organizacím možnost propojení GS s jejich lokální knihovnou, výsledky jsou ovšem vědeckým pracovníkům poskytovány pouze v rámci jejich působiště (akademická půda).

Současný příjemný vzhled GS vidíme na následujícím obrázku 2.1:



Obrázek 2.1

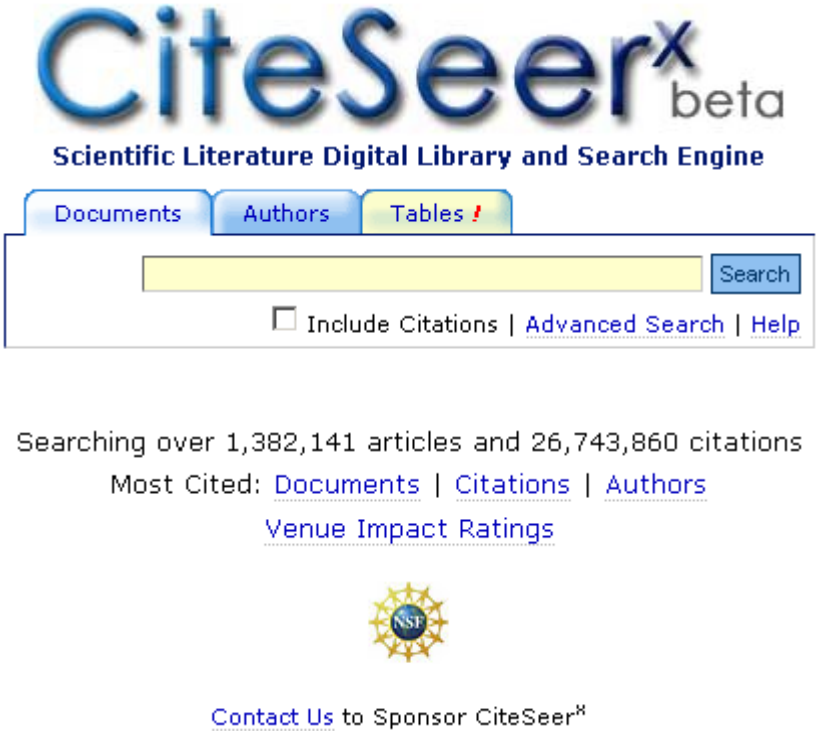
2.4.2 CiteSeer a CiteSeer^{x} beta

S projektem CiteSeer[1], dále už jen CS, a CiteSeerX[2], dále už jen CSX, je neodmyslitelně spojené jméno jeho zakladatele Prof. C. Lee Gilese, který se zasloužil se svým týmem o zrod celého projektu. Bylo to v roce 1997, kdy světlo světa spatřil projekt CS v NEC Research Institutu v New Jersey. Od roku 2003 přešla správa pod Pennsylvánskou státní univerzitu, konkrétně fakultu Informačních věd a technologií. Během deseti let CS narostl počet indexovaných dokumentů na úctyhodných 750 000 a přitom denně obsluhoval více než 1.5 milionu dotazů. Na základech CS staví nová verze CSX. Oba

1 http://citeseer.ist.psu.edu/

2 http://citeseerx.ist.psu.edu/

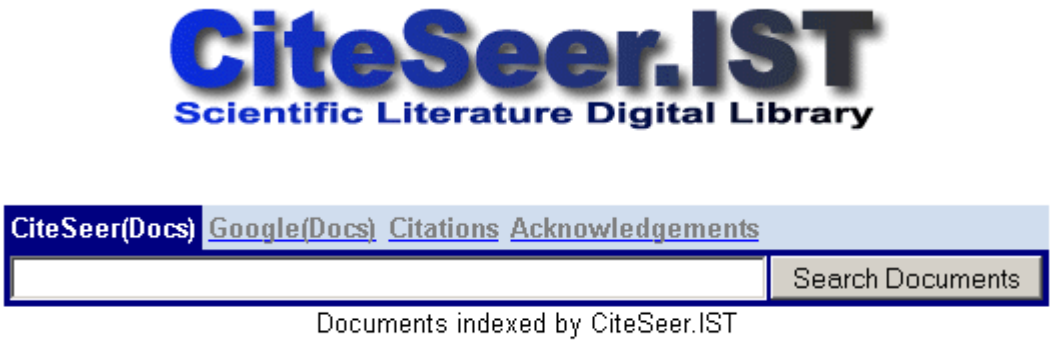
projekty jsou veřejné vyhledávací systémy a digitální knihovny pro vědu a akademické práce se zaměřením na počítače a obecně informační vědu. Původní CS byl historicky první, který uměl poskytovat automatizované citační indexy a citační odkazování používající metodu ACI (Autonomní citační indexování). S příchodem nové verze byl vývoj CS pozastaven a nyní ze svých stránek odkazuje na CSX. Za zmínku stojí, že CS sponzorovala NASA, Microsoft Research a National Science Foundation, který sponzoruje vývoj CSX doposud. CSX se pyšní novými metodami na zpracování metadat z PostScriptových[1] a PDF[2] výzkumných článků. CSX dále nabízí extrakce základních údajů, jako je název díla, výčet autorů, abstrakt, statistiky citování článku, s tím spojené citační extrakce, odkazy na obdobné články, odkazy na stažení indexovaného článku. CS dále na svých stránkách uvádí, že se snaží poskytovat nejnovější verze článků, sklízí nebo chceme-li prochází pravidelně indexované weby s články. U CSX je možnost se zaregistrovat a mít lepší možnost dostávat upozornění podobné RSS, spravovat vlastní kolekci článků aj. možnosti, které u Google Scholar nenalezneme.[10] Současný interface CSX a CS je vidět na obrázcích 2.2 a 2.3. Novější verze nabízí oproti staršímu CS možnost pokročilých hledání. Uživatel může hledat podle názvu dokumentů nebo jmen autorů.



obrázek 2.2

1 PostScript je programovací jazyk určený ke grafickému popisu tisknutelných dokumentů.

2 Portable Document Format – Přenosný formát dokumentů



obrázek 2.3

2.5 Skript Tomáše Lokaje

Skript Tomáše Lokaje, dále jen STL, byl vyvinut v jazyce Python na konci roku 2008 v rámci výzkumného projektu na FIT VUT v Brně. Vstupem STL jsou již převedené vědecké články na obyčejný text. Dále uváděný text je převzat z privátních stránek NLP WIKI FIT VUT v Brně[11] Skript se spouští s přepínačem '-f' pro jeden soubor a '-d' pro celou složku se soubory. Jako druhý parametr musí být pro STL uvedena cesta a název souboru (složky).

Příklad spuštění STL:

```
./clanky2meta.py -f text_txt/005.txt
./clanky2meta.py -d text_txt
```

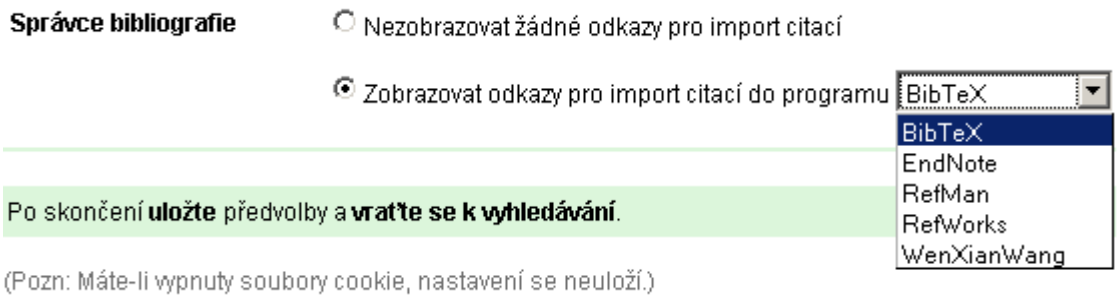
Dále autor uvádí, že program přijímá na vstupu pouze textové soubory s příponou 'txt' a jako výstup je 'xml' soubor obsahující extrahované meta informace. XML soubory se ukládají do složky 'xml' v pracovním adresáři. To není úplně pravda, dostupná verze se bohužel zbytečně zabývá extrakcí i jiných koncovek. Tento drobný nedostatek byl v této práci opraven. STL se snaží z převedeného dokumentu extrahovat autory, jejich emailové adresy, adresy jejich pracovního působiště, titulek, abstrakt a reference. STL pro extrakci využívá slovníky převzaté z diplomové práce Tomáše Petránka.

3 Analýza zvolených systémů

Tato kapitola vizuálně představuje webová rozhraní již uvedených vyhledávacích systémů. V obrázcích jsou názorně vyznačena a popsána důležitá nastavení a jednotlivé části výsledků hledání. Zabývá se možnostmi a způsoby dolování nebo chceme-li extrakcí klíčových metadat vhodných k následnému srovnávání se STL. Pro popis STL je zde věnována celá kapitola, která do větší hloubky přibližuje způsob práce skriptu. Na závěr je zde uveden příklad formátu výstupu STL

3.1 Google Scholar metadata

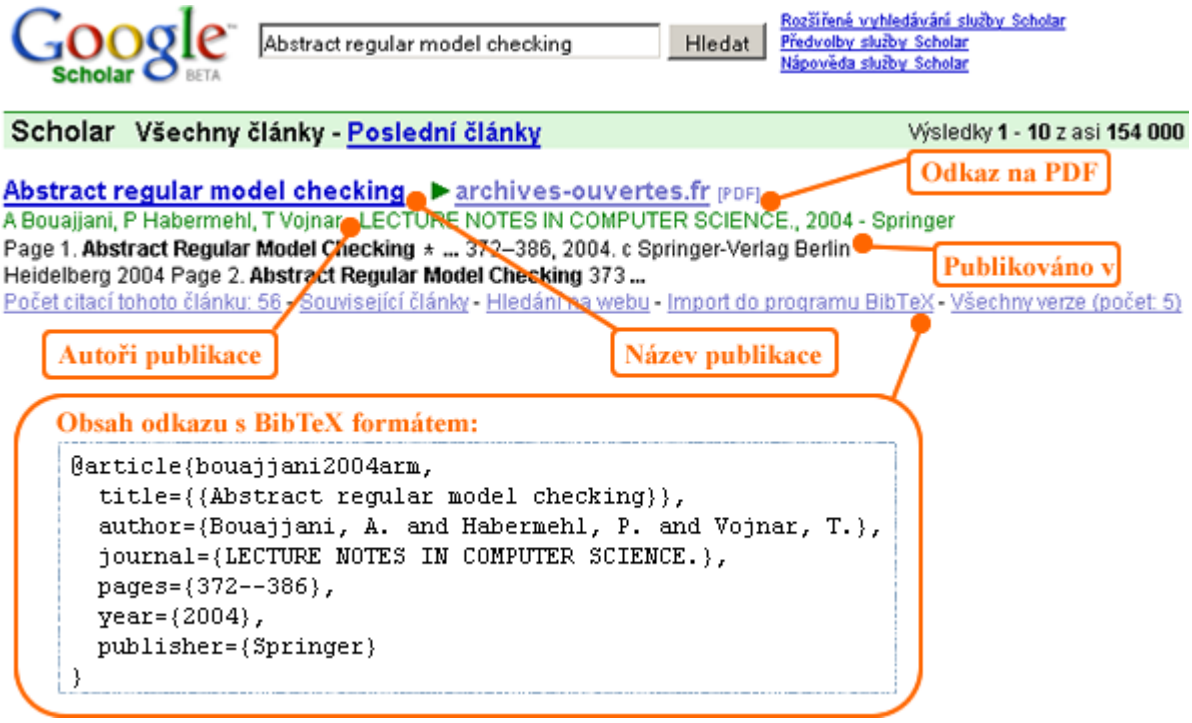
Vyhledávač GS nabízí v Předvolbách služby Scholar možnost zapnutí funkce zobrazování odkazu pro import citací do různých formátů. V nabízených formátech samozřejmě nechybí BibTeX.



Obrázek 3.1

Jak je vidět na obrázku 3.1, GS upozorňuje na nutnost používání cookie. U GS není dostupná možnost se přihlásit a spravovat si ostatní nastavení, proto jsou k zapamatování použity cookies[1], které se ukládají v prohlížeči uživatele GS.

1 Cookie v protokolu HTTP označuje malé množství dat, která WWW server pošle prohlížeči.

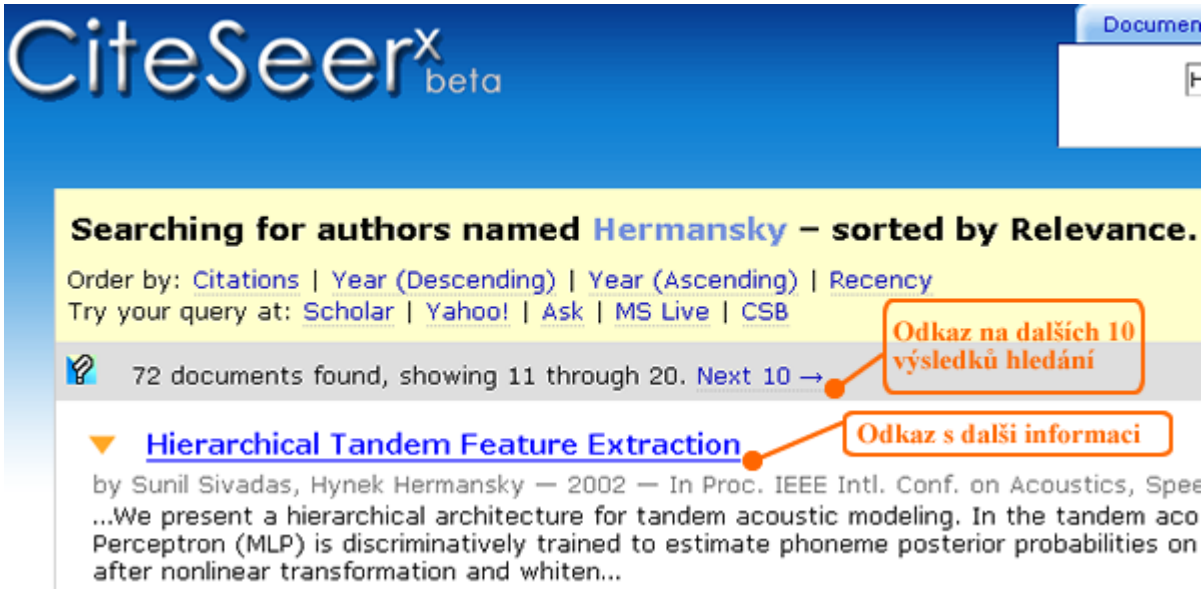


Obrázek 3.2

Na obrázku 3.2 vidíme popsány nejdůležitější části jednoho výsledku hledání na GS a obsah BibTeX importu, který by se nám po kliknutí běžně otevřel na nové stránce. Stránka s BibTeX importem je jen prostý text, tzn. není formátovaný HTML či XHMTL tagy. To je oproti předchozím informacím, které byly formátovány HTML tagy, velká výhoda.

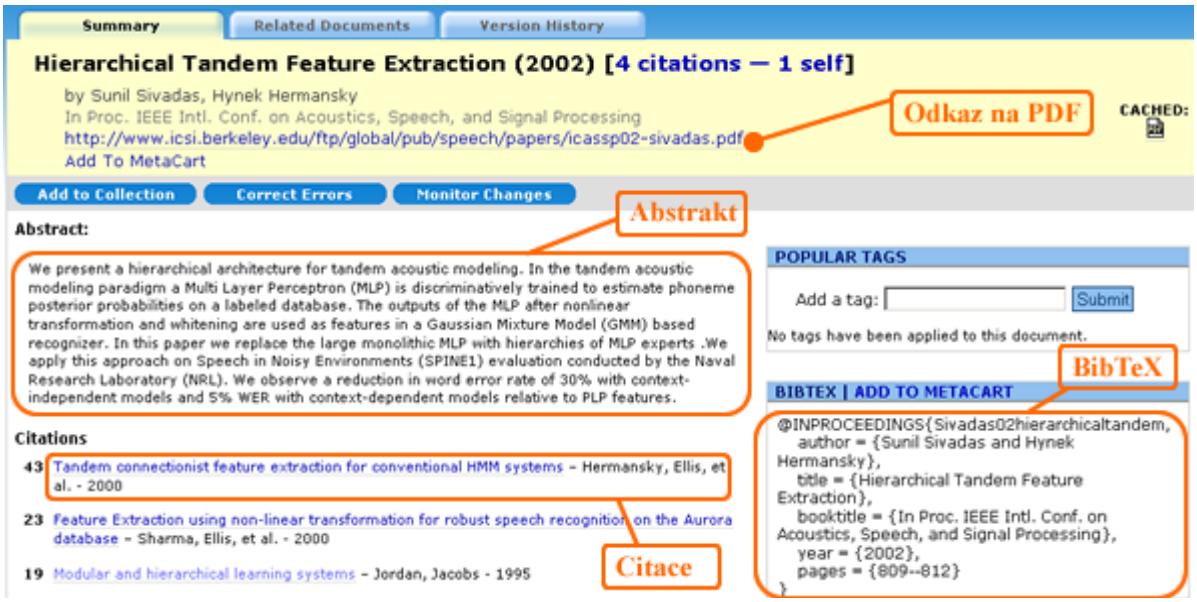
3.2 CiteSeer^x metadata

Po každém dotazu hledání CSX nabídne stránku s nejrelevantnějšími výsledky, jak vidíme na obrázku 3.3. Dotazy nemusejí být řazeny jen podle relevance, jak je ostatně na 3.3 vidět.



Obrázek 3.3

CSX narozdíl od GS nabízí možnost publikaci více prozkoumat. (GS odkazuje buď na dokument nebo na stránku zdroje, kde cenné informace sice jsou, ale nejsou v homogenním tvaru.) Na posledním obrázku 3.4 vidíme již konkrétní publikaci. Lze zde nalézt odkaz ke stažení publikace a BibTeX formát. Narozdíl od GS, je zde navíc uveden abstrakt publikace a uvedené citace. Číslo u citací indikuje, kolikrát byla stejná citace použita v jiných publikacích.



Obrázek 3.4

3.3 Extrakce metadat STL

Vstupem pro extrakci je textový soubor převedený z PDF souboru. Zde již vzniká možnost prvního zatížení chybou možným nekvalitním převodem. Skript používá slabou detekci špatně převedených souborů, která je založena na počtu špatných znaků. Je proto na místě, aby byl vstup lépe ověřován, a nedocházelo tak ke špatným extrakcím. STL nevhodné či nepoužitelné soubory zapisuje do souboru *error.log* ve výstupní složce.

STL si nejprve upraví vstupní text, aby byl bez diakritiky a špatně dekódovaných znaků. To vše pro následné použití slovníků na vyhledávání jmen. Následně hledá hlavičku, kde by měl nacházet název odborné práce, jména, adresy autorů a jejich emailové adresy. Hlavička zpravidla končívá před začátkem abstraktu. Hlavičky jsou v některých dokumentech formátovány do sloupců, pak jsou hlavičky uvedeny pod sebou. Taková hlavička se musí přeformátovat. Autor skriptu uvádí následující příklad, který je formátován do sloupců:

```
          Anup K. Sen                              Amitava Bagchi
Indian Institute of Management Calcutta      School of Management
   Joka, D. H. Road, Post Box 16757       University of Texas at Dallas
       Calcutta 700 027, INDIA                  Richardson, Texas 75083
        sen@iimcal.ac.in                         abagchi@utdallas.edu
```

A následně uvádí upravenou hlavičku, která už může být skriptem správně extrahována. Informace jsou zde uvedeny ve správném pořadí za sebou:

```
Anup K. Sen, Indian Institute of Management Calcutta, Joka, D. H.
Road, Post Box 16757,  Calcutta 700 027, INDIA, sen@iimcal.ac.in


Amitava Bagchi, School of Management, University of Texas at Dallas,
Richardson, Texas 75083, abagchi@utdallas.edu
```

STL se snaží najít v hlavičce dokumentu emailovou adresu autora. Pokud nalezne adresu složenou, rozkládá ji na dílčí adresy.

Příklad složené adresy:

```
{chelcicky,havlicek,adler,skocdopole}@stud.fit.vutbr.cz
```

Po extrakci jmen autorů jsou jim následně přiřazovány emailové adresy. Přiřazení se provádí podle nejvyššího počtu shodných podřetězců. K extrakci jmen autorů se používá již zmíněný slovník jmen, který obsahuje přibližně 22000 křestních jmen. Pro lepší přesnost určení jmen autorů jsou použity i jiné slovníky, například takové, které obsahují i názvy států. Pro extrakci adres pracovišť autorů a jejich následné přiřazení jsou použity další slovníky, které obsahují slova charakterizující adresu, to jsou názvy měst, států a názvy institucí. U extrakce titulků bývá detekce asi nejsnadnější, titulek je v drtivé většině ve vědeckých dokumentech na prvním řádku. Některé vstupní texty mají díky převodu titulek uveden na jiném místě, proto STL používá další heuristiky k detekci titulku. Abstrakt práce bývá opět samostatný odstavec, který je uvozen slovem *Abstract* nebo *Introduction*. Formátovaný abstrakt do sloupců je opět přeformátováván. Reference bývají uvozeny slovy *Bibliography* nebo *References. (N*utně poznamenat, že to bývá i poslední část dokumentu.) STL hledá reference oddělené oddělovačem v hranatých závorkách, číslicí s tečkou, římskou číslicí, číslicí v závorce. Dále autor uvádí, že takovými způsoby jsou reference rozdělovány málokdy. STL proto používá slovník křestních jmen k rozdělení referencí.

3.3.1 Výstup STL

STL vypisuje v průběhu své aktivity na standardní výstup průběh práce na daném dokumentu. Zpracované informace ukládá do souboru ve formátu XML a v kódování UTF-8[1].

STL vypisuje následující hlášení:

```
Nacitam slovniky.....   hotovo
Zpracovavam soubor 'dokumenty/E99-1001.txt':
        Upravuji vstupni text.....                  hotovo
        Hledam hlavicku.....                        hotovo
        Hledam emaily.....                          hotovo
        Hledam autory.....                          hotovo
        Prirazuji emaily.....                       hotovo
        Hledam odkazy.....                          hotovo
        Hledam titulek.....                         hotovo
        Hledam adresy.....                          hotovo
        Prirazuji adresy.....                       hotovo
        Hledam abstrakt.....                        hotovo
        Hledam reference.....                       hotovo
        Vytvarim XML soubor.....                    hotovo
Soubor 'dokumenty/E99-1001.txt' uspesne zpracovan za 0.34s
```

Příklad XML souboru s extrahovanými metadaty může vypadat následovně:

```
<?xml version="1.0" ?>
<document ref="008.txt">
     <head>
          <headline>
               Incremental Identification of Inflectional Typ es
          </headline>
          <authors>
               <author>
                    <name>
                         Petra Barg
                    </name>
```

1 UCS Transformation Format - způsob kódování řetězců znaků Unicode/UCS do sekvencí bajtů .

```
                    <email>
                        barg@ling.uni-duesseldorf.de
                    </email>
                    <address>
                        Heinrich-Heine-Universitt     Dusseldorf,
a ...
                    </address>
                </author>
                <author>
                    <name>
                        James Kilbury
                    </name>
                    <email>
                        kilbury@ling.uni-duesseldorf.de
                    </email>
                    <address>
                        Heinrich-Heine-Universitt     Dusseldorf,
a  ...
                    </address>
                </author>
            </authors>
        </head>
        <abstract>
            We present an approach to ...
        </abstract>
        <reference>
            Petra Barg and  Markus Walther. 1998. Processing unknown
words in ...
        </reference>
        <reference>
            Steven Bird and  Ewan Klein. 1994. Phonological analysis
in typed feature ...
        </reference>
        <reference>
            Michael  R.  Brent.  1991.  Automatic  acquisition  of
subcategorization ...
```

```
    </reference>
</document>
```

4 Návrh a implementace systému

Tato kapitola se zabývá návrhem a implementací porovnávacího systému pro vyhodnocení STL úspěšnosti v porovnání s již představenými webovými systémy GS a CSX. Jsou zde uvedeny údaje vhodné k extrakci a následnému porovnání. Údaje jsou oproti předchozím kapitolám zobrazeny i ve zdrojovém kódu, v jakém jsou webovými systémy nabízeny. Podrobně se kapitola zabývá algoritmem implementovaným k dolování údajů z GS, CSX a STL. Algoritmus systému je zde nejprve popisován metodou shora dolů a poté jsou jednotlivé části popisovány zvlášť. Taky proto je tento algoritmus pro větší názornost popsán jak graficky, tak i slovně.

Stejně jako byl uveden v předchozí kapitole výsledný extrakt ze STL, je zde prezentován i extrakt z implementovaného systému. Poslední část podkapitoly se zabývá srovnávací heuristikou při porovnávání jednotlivých extraktů GS, CSX a STL.

4.1 Dostupné metadata ke srovnání

V předchozích kapitolách byly vyznačeny informace, které jsou z uváděných systému dostupné, tedy aspoň v homogenním tvaru. Následující tabulka 4.1 nám poukazuje, které informace můžeme vzájemně porovnávat. Zatímco CSX nabízí v rámci zvolených metadat celé spektrum informací, tak GS ve dvou případech, konkrétně u referencí a u abstraktu, pokulhává. TSL z publikovaných dokumentů rok a journal či booktitle získat ani nemůže, protože se zde tyto informace ani nevyskytují, tedy alespoň ne v drtivé většině. Z tabulky lze dále vyčíst, že autory a názvy vědeckých prací můžeme porovnávat se všemi uváděnými systémy. V následujících podkapitolách jsou uvedeny způsoby získávání informace.

Metadata \ systémy	TLS	Google Scholar Beta	CiteSeerX
Autoři	ano	ano	ano
Název vědecké práce	ano	ano	ano
reference	ano	**NE**	ano
abstrakt	ano	**NE**	ano
journal/booktitle	**NE**	ano	ano
Rok vydání	**NE**	ano	ano

Tabulka 4.1

4.1.1 Získání autorů z CSX

U CS lze extrahovat autory z HTML kódu přímo pod názvem díla nebo z BibTeX formátu, jak je vidět na obrázku 3.4.

Pod názvem díla se nachází níže uvedený tag *div,* ve kterém jsou za anglickou předložkou *by* uvedení autoři publikace. Třída *char_increased char_indented char_mediumvalue padded* je na stránce unikátní. Oproti tomu BibTeXový formát je uveden v často se vyskytujícím *divu content*.

```
<div class="char_increased char_indented char_mediumvalue padded">
        by Sunil Sivadas, Hynek Hermansky
      </div>
```

4.1.2 Získání názvu a roku publikování vědecké práce z CSX

Název lze získat ze stránek obsahující výsledky hledání viz obrázek 3.3 nebo na stránce věnující se publikaci, viz 3.4. Zde je název uveden opět 2x, jednou jako nadpis a podruhé se nachází v BibTeX formátu. Jako nadpis je uveden v *h1* s třídou *primaryheader*. Za textem následuje, v závorkách uvedený, rok vydání vědecké publikace, ten lze zjistit i z BiBTeXových údajů.

```
<h1 class="primaryheader">

        <span class="Z3988" title="ctx_ver=Z39.88-
2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Akev%3Amtx
%3Ajournal&amp;rft_id=info:doi/10.1.1.38.4439&amp;rft_id=http%3A%2F
%2Fciteseerx.ist.psu.edu%2Fviewdoc%2Fsummary%3Fdoi
```

```
%3D10.1.1.38.4439&amp;rft.atitle=Spectral+Basis+Functions+From+Discriminan
t+Analysis&amp;rft.jtitle=INTERNATIONAL+CONFERENCE+ON+SPOKEN+LANGUAGE+PROC
ESSING&amp;rft.date=1998&amp;rft.genre=proceeding&amp;rft.au=And
%2C+Hynek+Hermansky&amp;rft.au=Hermansky%2C+Hynek&amp;rft.au=Malayath
%2C+Narendranath"> </span>

      Spectral Basis Functions From Discriminant Analysis
      (1998)

    </h1>
```

4.1.3 Získání referencí z CSX

Citace, které obsahuje publikace, jsou uvedeny v levé dolní části stránky. Ostatně, jak to opět můžeme vidět na obrázku 3.4. Název citace se nachází, jako odkaz s třídou *citation_only* na citovaný článek, uvnitř buňky tabulky s třídou *citelist*.

```
<table class="citelist">
        <tbody><tr><td>

          <span class="char_emphasized">590</span>
        </td>
        <td>

          <a class="citation_only" href="/showciting?cid=20815">Coloured
Petri Nets: Basic Concepts, Analysis Methods and Practical Use</a>
          - Jensen
           - 1996
        </td></tr>
```

4.1.4 Získání abstraktu z CSX

Abstrakt článku se nachází za nadpisem Abstrakt v párovém tagu *p* s třídou *para4*. Rozsáhlé abstrakty jsou kráceny a ukončeny třemi tečkami, jak je vidět na spodním příkladu, kde je text abstraktu výrazně zkrácen z pragmatických důvodu této práce.

```
<h2 class="topic_heading">Abstract:</h2>
  <p class="para4">: This article presents several different kinds of
Petri nets...</p>
```

4.1.5 Získání booktitlu či journalu z CSX

Tyto údaje se nacházejí pod autory nebo v BibTeXových informacích. Booktitle či journal je v divu s třídou *char_increased char_indented char6 padded*.

```
<div class="char_increased char_indented char6 padded">
        Proc. of ICASSP 00
      </div>
```

4.1.6 Získání publikace z GS

Soubor publikace lze získat z párového tagu *a* s parametry *href* a *title* obsahující stejnou adresu uvozenou uvnitř celého tagu *a*.

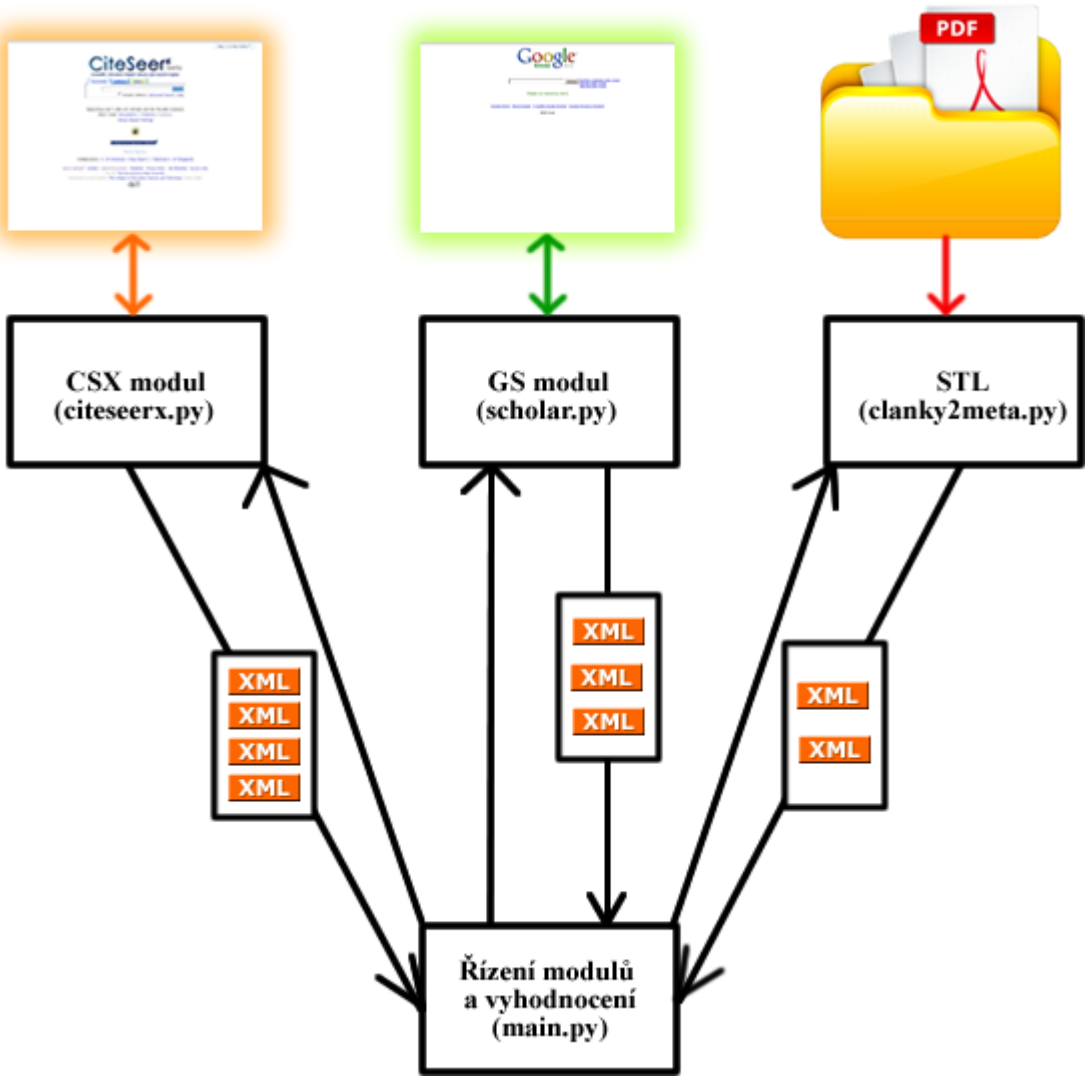
```
<a href="http://tmitwww.tm.tue.nl/staff/wvdaalst/Publications/p105.pdf"
title="http://tmitwww.tm.tue.nl/staff/wvdaalst/Publications/p105.pdf">http
://tmitwww.tm.tue.nl/staff/wvdaalst/Publications/p105.pdf</a>
```

4.1.7 Získání informací z GS

U GS je situace podstatně jednodušší. Na rozdíl od CSX jsou informace poskytovány čistě v BibTeX formátu bez HTML značek. Všechna podstatná metadata lze získat jednoduchým regulárním výrazem. Za připomenutí stojí, že ne vždy se zde nacházejí všechny důležité údaje k extrakci.

4.2 Algoritmus systému

Pro implementaci byl zvolen vysokoúrovňový a zároveň multiplatformní skriptovací jazyk Python verze 2. Celý systém je rozdělen do čtyř modulů. Algoritmus systému postupuje sériovým způsobem. Výsledky jednoho modulu se použijí pro vstup dalšího modulu. Na následujícím obrázku 4.1 máme zobrazen celý princip systému. Nejprve řídící modul požádá CSX modul, nechť sbírá informace z dotazů na CSX. Modul ukládá získané informace do složky *citeseerxsoubory*. Stáhne PDF dokument, který následně převede na obyčejný text. Pokud proběhne převod v pořádku, ukládá stažené informace do XML souboru. Jakmile CSX modul skončí, je spuštěn GS modul, který hledá informace o dokumentech podle názvu titulů z XML z adresáře *citeseerxsoubory*. GS modul opět ukládá získané informace do XML souborů, ale do složky *scholarsoubory*. Následně je spuštěn STL, který se snaží extrahovat informace z převedených PDF souborů. Nakonec probíhá statistické vyhodnocení.

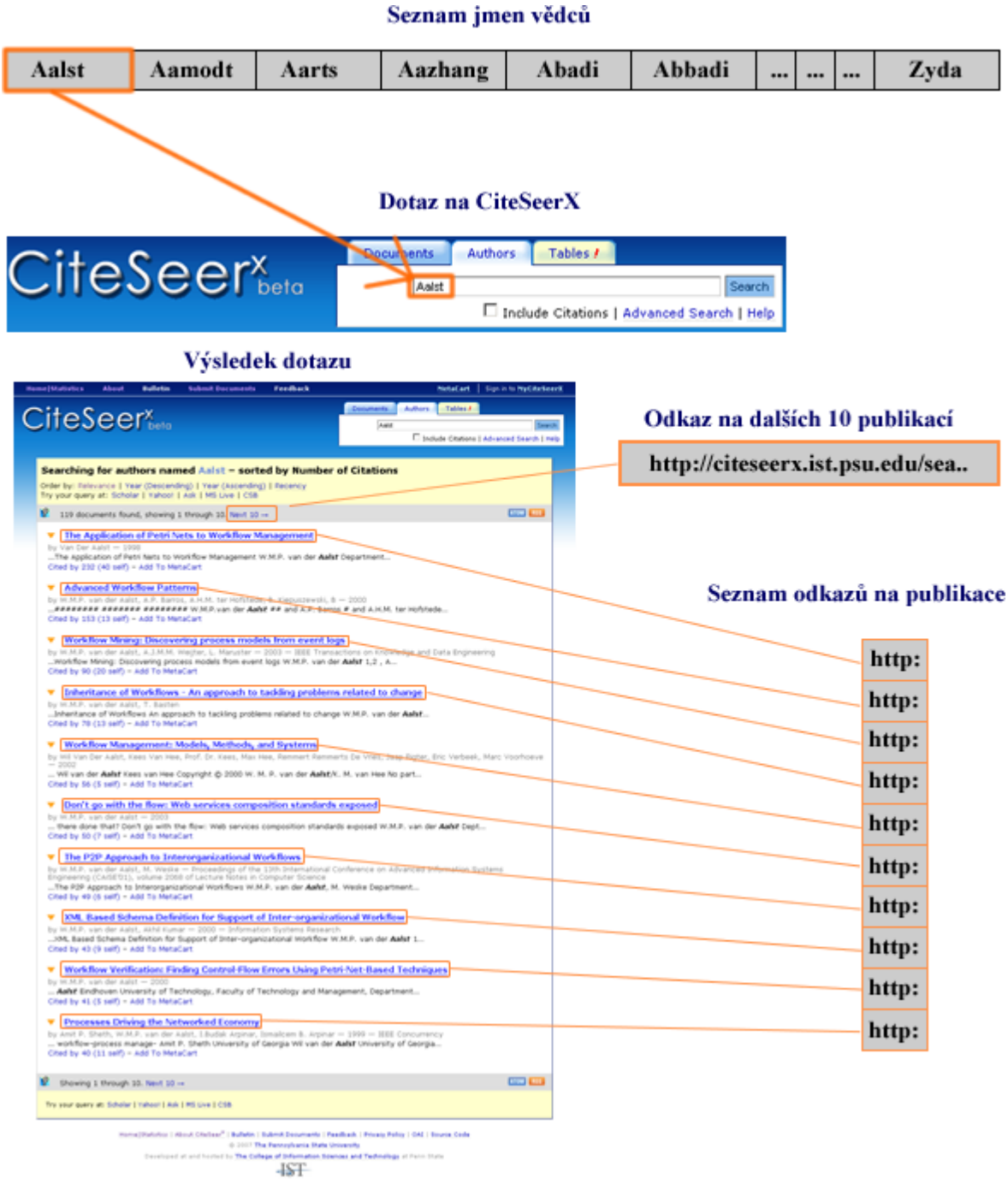


Obrázek 4.1

4.2.1 CSX modul

CSX na svých hlavních stránkách nabízí odkaz[12] na vlastní statistiky nejcitovanějších 10000 autorů. Z těchto autorů je vybráno cca 5500 autorů, kteří jsou unikátní, tzn., že už se v seznamu jejich jméno nevyskytuje. Následující obrázek 4.1 popisuje algoritmus zpracování stránek CSX. Nejprve se ze seznamu autorů vybere jeden autor. Zadá se dotaz na jeho jméno, tj. stáhneme stránku se seznamem (maximálně) deseti publikací od autora. Na této stránce se extrahují nalezené odkazy na stránky týkající se konkrétních publikací. Stránky v takovém seznamu jsou následně modulem stahovány. Extrakce dat ze stránky začíná teprve pokud: ze stejné url adresy nebyl už stahován, dokument ke stažení lze stáhnout, nachází se v PDF formátu a jeho převod na obyčejný text proběhl úspěšně. Převod i kontrolu provádí shellový skript *prevodPdf.sh* napsaný a upravený Robertem Kalmárem v rámci NLP projektu na FIT VUT v Brně. Tento skript provádí převod pomocí nástroje

pdftotext. Po extrakci všech stránek a následném uložení do souborů ve formátu XML se pokračuje na další stránku se seznamem dalších publikací a vše se takto opakuje, dokud jsou k dispozici další publikace. Název XML souboru je odvozen od názvu PDF souboru. V případě, že se vyskytne stejnojmenný soubor, je mezi původní název a koncovku ".xml" vloženo náhodně vygenerované číslo.


Seznam jmen vědců
Aalst
Aamodt
Aarts
Aazhang
Abadi
Abbadi
...
...
...
Zyda
Dotaz na CiteSeerX
CiteSeerx beta
Documents
Authors
Tables
Aalst
Search
Include Citations | Advanced Search | Help
Výsledek dotazu
Odkaz na dalších 10 publikací
http://citeseerx.ist.psu.edu/sea..
Seznam odkazů na publikace
http:
http:
http:
http:
http:
http:
http:
http:
http:
http:

Obrázek 4.2

4.2.2 GS Modul

Modul, který řídí celý systém, si z XML souborů z adresáře *citeseerxsoubory* vytvoří seznam. Postupně se tento seznam prochází. Z každého souboru se zjistí název publikace. Jako parametr GS modulu je předáván název publikace a název XML souboru. GS modul položí na GS dotaz s konkrétním názvem publikace. Je nutné společně s dotazem zasílat cookies a informace o typu prohlížeče. Je-li ve výsledku hledání shodný nebo velmi podobný název publikace, stáhne se BibTeX import. Z něho jsou extrahovány metadata o publikaci. V těchto informacích se mohou vyskytovat následující escape[poznamka] sekvence:

```
{\v{c}} {\'c} {\"a}
```

Tučné písma znamenají původní písmo bez diakritiky. Pokud vše proběhlo v pořádku jsou získané informace uloženy do stejnojmenného XML souboru jako u CSX modulu, ale do složky *scholarsoubory*.

4.3 Ukládání metadat

Struktura XML GS se od CSX takřka neliší. Ovšem oproti struktuře používané STL jsou zde nepatrné změny. Pro ukládání se používá *xml.dom.minidom* - odlehčená DOM implementace. *Journal* a *booktitle* jsou ukládány společně do tagu *journal-booktitle*.

Ukázka XML vytvářené CSX modulem:

```
<?xml version="1.0" encoding="utf-8"?>
<document ref="p3.pdf">
	<headline>
		Children As Digital Motion Picture Authors
	</headline>
	<url>
		http://www.dgp.toronto.edu/people/RMB/papers/p3.pdf
	</url>
	<authors>
		<name>
			Ronald Baecker
		</name>
		<name>
			Ilona Posner
		</name>
	</authors>
	<journal-booktitle>
		MAD: A Movie Authoring and Design System. Companion Proceedings to CHI'96
	</journal-booktitle>
	<abstract>
		this paper is that authoring and creating motion pictures is compelling to children and educationally valuable in a wide variety of ways. Until recently...
	</abstract>
	<year>
		1999
	</year>
```

<reference>
Situating Constructionism Papert 1991
</reference>
<reference>
Methodology Matters: Doing Research in the Behavioral and Social Sciences McGrath 1995
</reference>
<reference>
Technologies for Knowledge-building Discourse Scardamalia, Bereiter 1993
</reference>
<reference>
Finding Art's Place: Experiments in contemporary education and culture Paley 1995
</reference>
<reference>
Computer Support for Authoring Motion Pictures Rosenthal 1995
</reference>
<reference>
in press). MAD: A Movie Authoring and Cohen, Friedlander, et al. 1996
</reference>
<reference>
New technologies, new literacies, new problems Reilly 1996
</reference>
<reference>
Watching Media Learning Buckingham 1990
</reference>
<reference>
Composing with Images: A Study of High School Video Producers Reilly 1994
</reference>
</document>

Ukázka XML vytvářené GSX modulem:

```
<?xml version="1.0" encoding="utf-8"?>
<document ref="p3.xml">
      <headline>
            Children as digital motion picture authors
      </headline>
      <authors>
            <name>
                  Baecker, R.
            </name>
            <name>
                  Posner, I.
            </name>
      </authors>
      <journal-booktitle>
            The Design of Childrens Technology
      </journal-booktitle>
      <year>
            1999
      </year>
</document>
```

4.4 Srovnávací metody a statistické údaje

Za srovnávací metodu budeme považovat funkci, která nám porovná dva zadané řetězce znaků a určí nakolik jsou si podobné či shodné.

4.4.1 Srovnávací metody u GS modulu

Už modul GS používá dvě srovnávací metody. Srovnáváme zde hledané dílo s nalezeným prvním čili většinou nejrelevantnějším výsledkem. První metoda je na absolutní shodu. Ta počítá kolik procent odpovídalo shodě n-tého slova prvního řetězce n-tému slovu druhého řetězce. Jiné než alfanumerické znaky jsou ignorovány, to samé platí i o národních znacích.

Například následující dva řetězce jsou shodné na 80%.

```
"Verifying properties of parallel programs"
"Verifying properties of serial programs"
```

Druhá metoda u GS modelu se aplikuje na předchozí řetězec pokud nedosáhl shody alespoň na 70 %. Tato metoda už neporovnává n-té slovo prvního řetězce s n-tým slovem druhého řetězce, ale každé slovo s každým. Pokud se u druhé metody dosáhne alespoň 65% úspěšnosti, jsou vyhodnoceny řetězce jako podobné a metadata nalezeného odkazu jsou extrahovány. Mohlo by se zdát, že druhou metodou se zavedou chyby, které by mohly výrazně zatížit chybou statistiky vyhodnocení, ale právě díky GS dostáváme zpětnou vazbu s CSX. Pokud CSX titulek vyhodnotil špatně, najde se u GS hodně podobný, dost pravděpodobně ten správný. Pokud jsou oba odlišné tak potom záleží na tom, jak dopadne STL extrakce.

4.4.2 Srovnávání výsledků

Srovnávání výsledků uložených v XML probíhá taktéž elegantní metodou. Nejdříve se hledají stejnojmenné soubory ve složkách s XML od GS modulu (*scholarsoubory*), od CSX modulu (*citeseerxsoubory*) a STL (*xml*).

Výsledek participující k jednomu dokumentu se zde porovnává každý s každým. Funkce zde parametricky rozlišuje, jestli se v řetězci mohou vyskytovat iniciály autorů. Dost často se stává, že někde jsou jména autorů napsána celým jménem a někdy jen částí, typicky křestní jméno. Jak taková funkce pracuje? Nejdřív se v řetězci převedou velká písmena na malá, rozdělí se řetězce na slova, jako oddělovače bereme opět všechny nealfanumerické znaky včetně národních znaků. Dostaneme dva seznamy slov, obsah takových seznamů seřadíme podle délky slov – nejkratší slova typicky iniciály jsou na konci seznamu. Poté delší seznam takto seřazených slov je procházen a kontroluje se v menším či stejně dlouhém seznamu, jestli se zde nachází stejné slovo. Jestli je nalezeno, odebere se z krátkého seznamu a je připočítána shoda. Pokud je funkce volána s parametrem pro iniciály, procházejí se seznamy ještě jednou stejným způsobem, až na to, že jsou porovnávána zbylá slova o délce jedna s ostatními slovy - porovnáváme iniciály s celými jmény. Parametr pro iniciály nastavujeme pouze pokud porovnáváme autory a reference.

Příklad porovnání autorů:

Vitor Santos, Jose Castro, M. Isabel Ribeiro	R. Santos, Jose Castro, m. i. Ribeiro
vitor santos, jose castro, m. isabel ribeiro	r. santos, jose castro, m. i. ribeiro
['vitor', 'sanros', 'jose', 'castro', 'm', 'isabel', 'ribeiro']	['r', 'santos', 'jose', 'castro', 'm', 'i', 'ribeiro']
['ribeiro', 'castro', 'sanros', 'isabel', 'vitor', 'jose', 'm']	['ribeiro', 'santos', 'castro', 'jose', 'r', 'm', 'i']
['isabel', 'vitor']	['r', 'i']
['vitor']	['r']

Shoda se vypočítá podle vzorce 4.1. V našem případě by výpočet vypadal následujícím způsobem:
((6 * 2) / 14) * 100 = 85,714%

$$Shoda = \frac{poměr\ shodných\ dvojic \times 2}{počet\ všech\ slov} \times 100$$

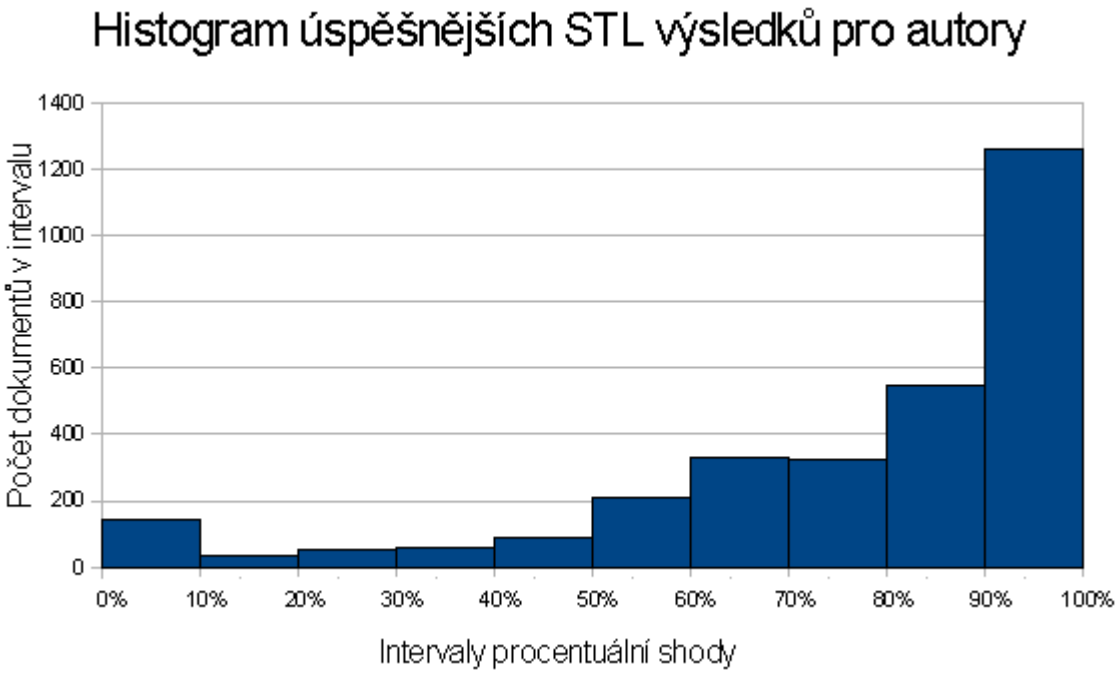
Vzorec 4.1

5 Vyhodnocení statistik

V této kapitole jsou zobrazeny histogramy ke každé vyhodnocované části. I když systém není typicky uživatelský a je jednoúčelový, zpracovává na standardní výstup histogramy s intervaly 10 a 20 % a další informace o průměrných hodnotách. V následujících podkapitolách jsou zobrazeny histogramy pouze s 10% intervaly. Všechny statistiky v konečném hodnocení pracují s 3045 publikacemi.

5.1 Vyhodnocení autorů

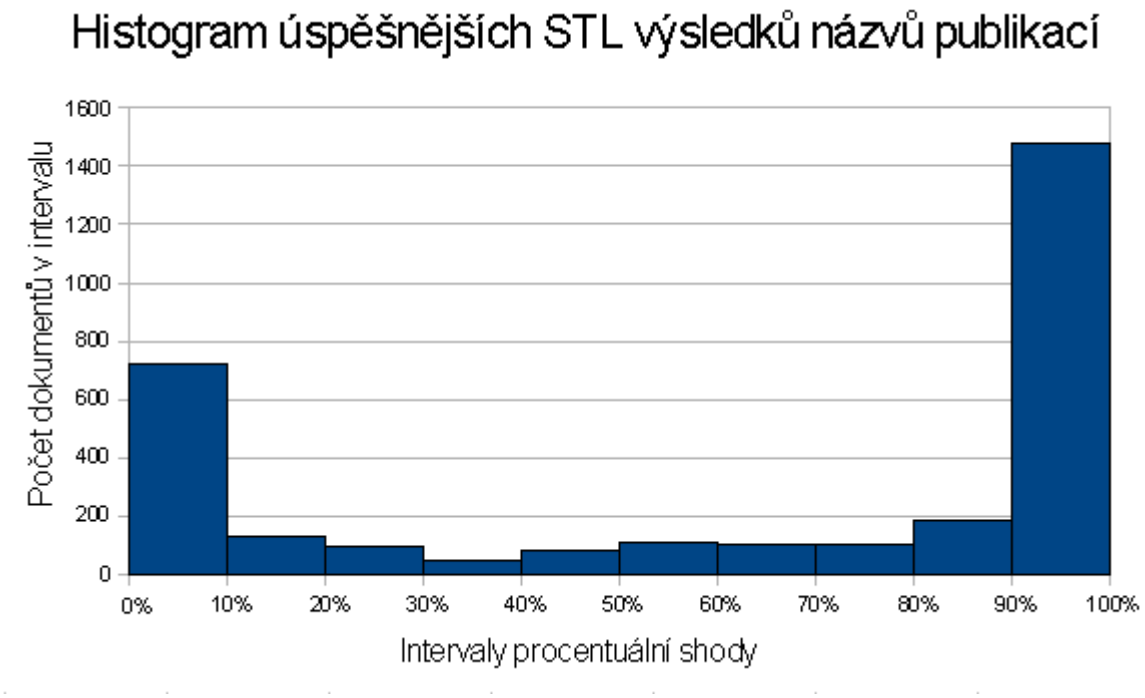
Průměr dosažených lepších hodnot STL byl 77 % v porovnání s GS a CSX. Medián lepších hodnot je 84.2 %. 32.2 % extrahovaných autorů bylo 100% shodných s GS nebo s CSX. Pouze 4.37 % se absolutně neshodovalo s GS nebo CSX extrakci. 30.15 % výsledků STL bylo lepších než srovnání CSX s GS. CSX a GS se shodovaly v 50 % případů.



Histogram 5.1

5.2 Vyhodnocení názvu vědecké práce

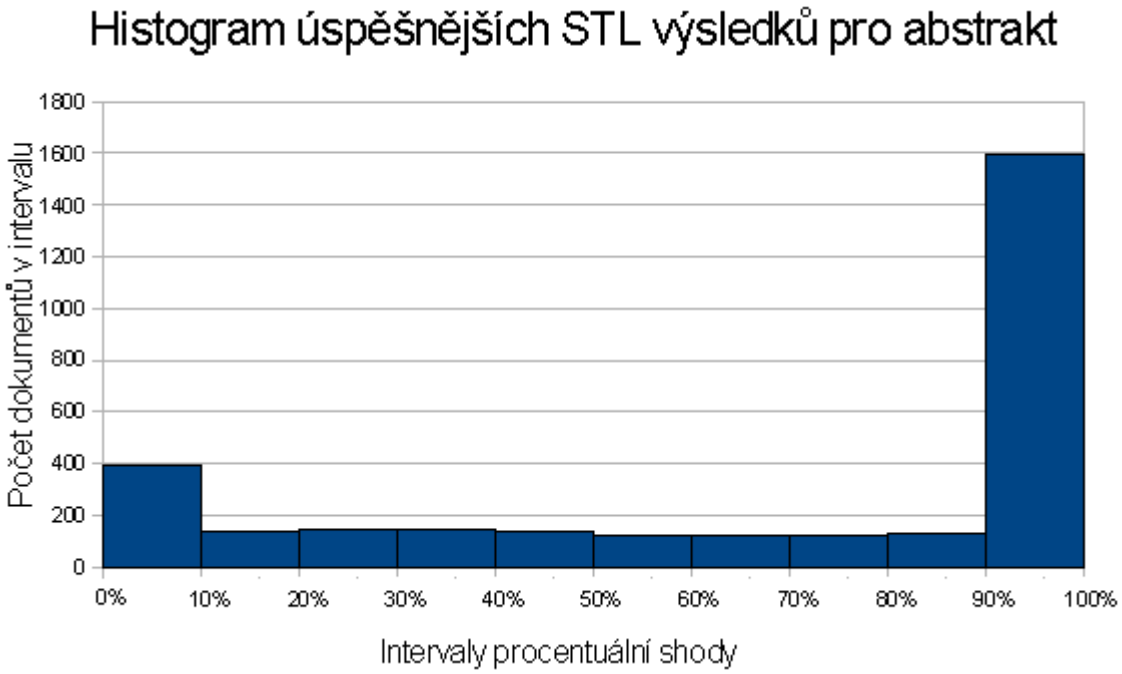
Průměr dosažených lepších hodnot STL byl 63.3 % v porovnání s GS a CSX. Medián lepších hodnot je 87.5 %. 43.7 % extrahovaných názvů bylo 100 % shodných s GS nebo s CSX. Až 20 % se absolutně neshodovalo s GS nebo CSX extrakci. Pouze 7 % výsledků STL bylo lepších než srovnání CSX s GS. S tím hodně souvisí, že CSX a GS se shodovaly v 85.6 % případů.



Histogram 5.2

5.3 Vyhodnocení abstraktů

Vyhodnocených 3045 abstraktů dosahovalo průměrně 68.7 % úspěšnost. Medián se pohybuje okolo 92.75 %!

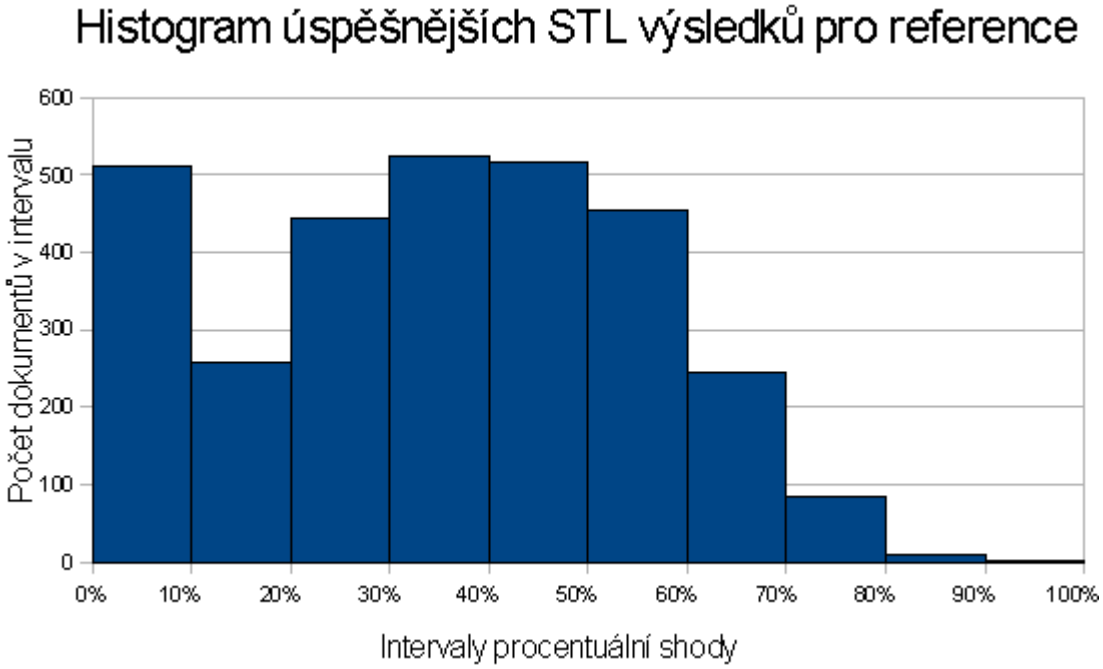


Histogram 5.3

5.4 Vyhodnocení referencí

Průměr dosahovaných hodnot je pouhých 34.5 % Medián se moc neliší, má 36.3 %.

Za zmínku stojí, že do intervalu 90 % - 100 % se dostalo pouze jedno jediné vyhodnocení referencí.



Histogram 5.4

5.5 Porovnání se statistikami Tomáše Lokaje

Tomáš Lokaj provedl malé ruční srovnání 50 článků ze serveru ArXiv[1]. Následující tabulka 5.1 prezentuje Lokajovy a naše dosažené výsledky:

	STL 50 článků	**3045 článků**
Titulek	76,00%	0,00%
Autoři	72,00%	77,00%
Abstrakt	64,00%	68.67%
Reference	90,00%	34.57%

Tabulka 5.1

1 http://arxiv.org/

5.6 Celkové zhodnocení statistik

Jak je vidět z tabulky 5.1 výsledky u autorů a abstraktu článku jsou si hodně podobné. Titulky se liší o 12%. Největší propad lze zaznamenat u referenci článků. Problém u referencí spočívá v rozdílu, co je ještě reference. CSX sem zařazuje rok publikování článku a STL číslo reference. Další zatížení chybou může nastat u kódování národních znaků a nekvalitním převodem z PDF formátu.

Při vyhodnocování takovéhoto objemu dat STL vykazoval programové nedostatky v podobě neošetřených podmínek.

6 Závěr

Vyhotovený systém potvrdil již předem známou domněnku, že na skriptu Tomáše Lokaje je stále co zlepšovat. Většina získaných statistických údajů se hodně podobala malé statistice Tomáše Lokaje. Vědecké publikace se liší především v drobných zásadách psaní, které v důsledku představují pro extrakci velké problémy. Lokaj tedy na stránce[11] projektu vhodně poznamenává, že je na skriptu stále nutné doplňovat různé šablony vědeckých prací. I když se výsledky nejeví na první pohled příznivě, je nutné poznamenat, že stejně tak, jak je Lokajův skript zatížen různými chybami, je i náš systém ne zcela dokonalý. Autor vidí nedostatky ve zpracování národních znaků a u referencí, které dopadly nejhůř, protože jsou k porovnávání poskytovaný ne zcela homogenní informace. Minimálně na těchto drobnostech by se v následujícím vývoji mělo zapracovat. Pozitivní je, že tato práce může poskytnout dalším studentům jednoduchý návod na zpracovávání a porovnávání informací.

Literatura

[1] *Metadata Management* [online]. 2009 [cit. 2009-05-25]. Dostupný z WWW: <http://www.profinit.eu/cz/it-reseni/data-management/metadata-management>.

[2] *BibTeX : Wikipedia, the free encyclopedia* [online]. 2003 , 23.5.2009 [cit. 2009-05-25]. Dostupný z WWW: <http://en.wikipedia.org/wiki/BibTeX>.

[3] BibTeX : Wikipedie, otevřená encyklopedie [online]. 2006 , 2.1.2009 [cit. 2009-05-25]. Dostupný z WWW: <http://cs.wikipedia.org/wiki/BibTeX>.

[4] *About BibTeX* [online]. 2006 [cit. 2009-05-25]. Dostupný z WWW: <http://www.bibtex.org/About>.

[5] VOLTR, J.. *Odborný článek - jak na něj* [online]. [2008] [cit. 2009-05-25]. Dostupný z WWW: <http://fyzsem.fjfi.cvut.cz/2008-2009/Leto09/html/publish/paper.html>.

[6] *Věda a publikační činnost doktorandů* [online]. Brno : Provozně ekonomická fakulta MZLU, 2009 , 24. 04. 2009 [cit. 2009-05-25]. Dostupný z WWW: <http://www.pef.mendelu.cz/cz/studium/doktorske/elektronicky_pruvodce/veda_publikace>.

[7] Academic Search Engines. *ONLINE SEARCHING FOR PROFESSIONAL OR ACADEMIC PURPOSES – PART II* [online]. 2007 [cit. 2009-05-25]. Dostupný z WWW: <http://www.lluiscodina.com/AcademicSearchEngines.ppt >.

[8] *O službě Google Scholar* [online]. 2009 [cit. 2009-05-25]. Dostupný z WWW: <http://scholar.google.cz/scholar/about.html>.

[9] *Google Scholar se rozvíjí - má Scirus konkurenci?* [online]. 2005 [cit. 2009-05-25]. Dostupný z WWW: <http://www.ikaros.cz/node/1871>.

[10] *About MyCiteSeer* [online]. 2007 [cit. 2009-05-25]. Dostupný z WWW: <http://citeseerx.ist.psu.edu/about/myciteseer>.

[11] https://merlin.fit.vutbr.cz/nlp-wiki/index.php/P%C5%99evod_metainformac%C3%AD

[12] http://citeseerx.ist.psu.edu/stats/authors?all=true

Seznam příloh

Příloha 1. DVD